Capítulo 3

Análisis de Regresión Simple

1. Introducción

El análisis de regresión lineal, en general, nos permite obtener una función lineal de una o más variables independientes o predictoras (X_1, X_2, ... X_K) a partir de la cual explicar o predecir el valor de una variable dependiente o criterio (Y). En el análisis de regresión lineal podemos diferenciar entre análisis de regresión lineal simple y análisis de regresión lineal múltiple. En el primero, se intenta explicar o predecir la variable dependiente Y a partir de una única variable independiente, X_1; mientras que en el segundo, contamos con un conjunto de variables independientes, X_1, X_2, ... X_K, para estimar la variable dependiente Y. En ambos casos, tanto la variable dependiente como la/s independiente/s están medidas en escala de intervalo o de razón.

En este capítulo nos vamos a ceñir al análisis de regresión lineal simple posponiendo para el próximo capítulo la regresión lineal múltiple que, como tendremos ocasión de apreciar, comparte mucho de lo que en estas líneas se recoge. El análisis de regresión lineal simple tiene por finalidad predecir y/o estimar los valores de la variable dependiente a partir de la obtención de la función lineal de la variable independiente. La anotación matemática de la ecuación de regresión simple se anota como sigue:

$$Y = a + b_{1x1} + e$$

ó

$$\text{presente} = a + b1_{\text{pasado}} + e$$

en donde:

Y es la variable a predecir;

a y b_{1x1} son parámetros desconocidos a estimar;

y e es el error que cometemos en la predicción de los parámetros.

No obstante, antes de proceder a la estimación de los parámetros, y con ellos a la concreción de una ecuación predictiva, debemos corroborar que, efectivamente, los datos sometidos a

análisis se adaptan a un modelo de regresión lineal. La lección la hemos estructurado en los siguientes puntos:

1. Exposición de los estadísticos que nos permiten valoración de la bondad de ajuste de los datos al modelo de regresión lineal simple.
2. Si los estadísticos certifican que entre los datos se produce una asociación lineal, podremos pasar a estimar los parámetros de la ecuación lineal (B_0 y B_1), a partir de los cuales podremos efectuar predicciones de la variable dependiente. Cabe advertir que en el supuesto caso en el que los estadísticos rechazaran la asociación lineal entre los datos, no significa que entre ellos se produzca otro tipo de relación (como la curvilínea).
3. Por último, exponemos la secuencia de pasos que nos permiten determinar lo arriba apuntado. En el análisis de regresión simple, y con la finalidad de obtener la mayor información posible respecto a la relación y asociación entre las dos variables, vamos a trabajar con tres Cuadros de Diálogos, a saber: Cuadro de Diálogo de Correlaciones Bivariadas. Cuadro de Diálogo de Gráficos; y Cuadro de Diálogo del Análisis de Regresión Lineal Múltiple.

2. Bondad de ajuste de los datos al modelo de regresión lineal simple

Antes de poder aplicar el modelo de regresión lineal simple para predecir los valores que alcanzará una determinada variable criterio, debemos certificar que los datos a los que sometemos a dicho análisis se ajustan al modelo de regresión lineal simple; o lo que es lo mismo, debemos analizar el grado de asociación lineal entre la variable dependiente y la independiente así como determinar la proporción de variabilidad de la variable dependiente explicada por la independiente.

Los principales estadísticos y pruebas que nos permiten valora la bondad de ajuste de los datos al modelo de regresión lineal simple son.

1.- Coeficiente de Correlación Lineal Simple (r).

Mide el grado de asociación lineal entre dos variables. Este estadístico oscila entre 1 (fuerte asociación lineal positiva: a medi-

da que aumenten los valores de una variable aumentarán los de la otra) y –1 (fuerte asociación lineal negativa: a medida que aumenten los valores de una variable disminuyen los de la otra). Cuando los valores de este estadístico se aproximen a 0 nos estará indicando que entre las dos variables no existe asociación lineal y, en consecuencia, carece de sentido determinar el modelo y/o ecuación de regresión lineal. Resulta muy interesante comparar este coeficiente junto con el Scatter Plot de la nube de puntos (*gráfico 1 del anexo de resultados*), ya que el gráfico nos ofrece una representación elocuente de la distribución y relación de las dos variables relacionadas. Si la nube de puntos forma una forma indefinida y muy dispersa, nos indica la inexistencia de relación entre las variables. Si por el contrario, se observa una forma definida y proximidad entre los puntos, habrá relación entre las variables caracterizada por la forma y distribución que adopte.

Para determinar si la asociación es estadísticamente significativa podemos contrastar la H_0 de que el coeficiente de correlación lineal es igual a 0; o lo que es lo mismo, que las dos variables están incorrelacionadas. Si el *p-valor* asociado al estadístico de contraste (r) es menor que el nivel de significación elegido (normalmente 0.05) rechazaremos H_0. En la matriz de correlaciones se recogen estos dos valores: en primer lugar aparece el grado de relación (r) que se produce entre las dos variables que cruzamos; y en segundo lugar, la significación estadística de esa relación.

Debemos hacer notar que pese a que estemos efectuando un análisis de regresión lineal bivariado, el proceso que seguimos es el del análisis de regresión multivariable. El cuadro de diálogo del análisis multivariado ofrece una información más rica de ahí la tendencia generalizada a utilizar éste en detrimento del cuadro de diálogo de regresión simple. Por esta razón, vamos a ver como en las salidas del ordenador, y pese a estar realizando un análisis con dos variables, a este coeficiente se le denomina coeficiente de Correlación Múltiple (Multiple R), residiendo la explicación en el hecho de que va a ser siempre el análisis multivariable el que apliquemos indistintamente si nos encontramos trabajando con dos variables, como es ahora el caso, o con más variables, como se verá en el próximo capítulo. No debemos confundir el coeficiente de correlación múltiple (mide el grado de asociación entre la variable dependiente y un conjunto de variables independientes), del los coeficientes

de correlación lineal simple (aparecen en la matriz de correlaciones).

2.- Coeficiente de Correlación Múltiple al Cuadrado o Coeficiente de Determinación (R Square "R^2").

El coeficiente de determinación se define a partir del coeficiente de correlación múltiple (R) y mide la proporción de variabilidad de la variable dependiente explicada por la variable independiente introducida o por la recta de regresión. Si el valor que resulta lo multiplicamos por 100, obtendremos el porcentaje de variabilidad explicada.

3.- Coeficiente de Determinación Ajustado (Adjusted R Square).

Pese a que R^2 se viene utilizando como medida de ajuste al modelo, presenta el inconveniente de que a medida que vamos incrementando el número de variables que participan en el modelo (será el caso propio del análisis multivariable) mayor es su valor de ahí que la R^2 sobrestime el verdadero R de la población. Por esta razón, algunos autores recomiendan utilizar el Coeficiente de Determinación Ajustado pues éste no aumenta, necesariamente, a medida que añadimos variables a la ecuación. Este estadístico queda ajustado por el número de observaciones y el número de variables independientes incluidas en la ecuación.

4.- Error Típico de Predicción (ETB).

El error típico de la predicción es la parte de la variable dependiente que dejamos de explicar ya sea porque nos falte alguna variable por introducir, o bien, porque las variables que hemos elegido no son más las adecuadas. Su cálculo se establece a partir de la desviación típica de la variable dependiente y el coeficiente de determinación ajustado.

5.- Análisis de Varianza.

La tabla de análisis de varianza que incluye en su salida el SPSS nos permite valorar hasta qué punto es adecuado el modelo de regresión lineal para estimar los valores de la variable dependiente. La tabla de análisis de varianza se basa en que la variabilidad total de la muestra puede descomponerse entre la variabilidad explicada por la regresión y la variabilidad residual. La tabla de ANOVA proporciona el estadístico F a partir del cual podemos contrastar la H_0 de que R^2 es igual a 0, la pendiente de la recta de regresión es igual a 0, o lo que es lo mismo, la

hipótesis de que las dos variables están incorrelacionadas. Si el *p-valor* asociado al estadístico F es menor que el nivel de significación (normalmente 0.05), rechazaremos la hipótesis nula planteada.

6.- Análisis de Residuales.

Como ya hemos comentado los residuos, "***e***", son la estimación de los verdaderos errores. En regresión lineal la distribución de la variable formada por los residuos debe ser Normal, esto es, los residuos observados y los esperados bajo hipótesis de distribución normal deben ser parecidos. Además, los residuos deben ser independientes. En consecuencia, el análisis de los residuales nos va a permitir no solo profundizar en la relación que se produce entre las dos variables, sino también, ponderar la bondad de ajuste de la regresión obtenida.

Para contrastar la supuesta normalidad de los residuales podemos recurrir, fundamentalmente, a la representación de dos gráficos: (1) ***el gráfico de residuales tipificados*** (*gráfico 2 del anexo de resultados*) nos da idea de cómo se distribuyen los residuos en relación a la distribución normal (que sería la que cabría esperar de los mismos). Si ambas distribuciones son iguales (la distribución de los residuos es normal) los puntos se sitúan sobre la diagonal del gráfico. Por lo contrario, en la medida que aparecen dispersos y formando líneas horizontales respecto a la diagonal, habrá más residuos y el ajuste será peor; (2) ***el gráfico de probabilidad normal*** (*gráfico 3 del anexo de resultados*) compara gráficamente, al superponer la curva de distribución normal, la función de distribuciones acumulada observadas en la muestra con la función de distribución acumulada esperada bajo supuestos de normalidad.

Por su parte el estadístico de ***Durbin-Watson*** mide el grado de autocorrelación entre el residuo correspondiente a cada observación y el anterior (si los residuos son independientes, el valor observado en una variable para un individuo no debe estar influenciado en ningún sentido por los valores de esta variable observados en otro individuo). Si el valor del estadístico es próximo a ***2*** los residuos están incorrelacionados; si se aproxima a ***4***, estarán negativamente incorrelacionados; y si se aproximan a ***0*** estarán positivamente incorrelacionados.

3. Estimación de los parámetros o coeficientes de regresión: la ecuación de predicción o ecuación de regresión simple

Una vez que ya hemos analizado el carácter e intensidad de la relación entre las variables, podemos proceder a estimar los parámetros de la ecuación de predicción o de regresión lineal. El criterio para obtener los coeficientes de regresión B_0 y B_1 es el de ***mínimos cuadrados.*** Este consiste en minimizar la suma de los cuadrados de los residuos de tal manera que la recta de regresión que definamos es la que más se acerca a la nube de puntos observados y, en consecuencia, la que mejor los representa.

Los estadísticos asociados a la variable independiente que a pasado a formar parte del modelo de regresión simple son:

1.- Coeficiente de regresión B.
Este coeficiente nos indica el número de unidades que aumentará la variable dependiente o criterio por cada unidad que aumente la variable independiente.

2.- SEB.
Error típico de B.

3.- Coeficiente Beta.
El coeficiente Beta es el coeficiente de regresión estandarizado. Expresa la pendiente de la recta de regresión en el caso de que todas las variables estén transformadas en puntuaciones Z.

4.- Constante.
El valor de la constante coincide con el punto en el que la recta de regresión corta el eje de ordenadas. En la ecuación de predicción se mantiene constante para todos los individuos. Cuando las variables han sido estandarizadas (puntuaciones Z) o si se utilizan los coeficientes Beta, la constante es igual a 0 por lo que no se incluye en la ecuación de predicción.

5.- Tolerancia.
Tolerancia es la proporción de variabilidad no explicada por el resto de variables ($1-R^2$). Cuanto mayor sea la T más independiente es la variable en cuestión.

6.- Valor T.

El estadístico T nos permite comprobar si la regresión entre una variable independiente y la dependiente es significativa. Si el *p-valor* asociado al estadístico T (Sig T) es mayor al nivel de significación (normalmente 0.05) rechazaremos que la regresión sea significativa para las dos variables relacionadas.

En nuestro caso la significación del estadístico T asociado al modelo generado con la única variable independiente que disponemos es inferior a 0.05 de ahí que podamos ratificar el carácter predictivo de dicha variable y podamos, en consecuencia, exponer la ecuación del modelo. En el ejemplo que recogemos en la sección de Resultados, la transcripción de los resultados a la ecuación quedaría como sigue:

$$Y = a + b_{1x1} + e$$

ó

$$\text{presente (p7A)} = 0{,}51 + 0{,}87_{\text{pasado (p7B)}} + e$$

en el supuesto caso de que los valores de las variables siguieran una escala diferente, tendríamos que estandarizar utilizando los coeficientes Beta, y no B. Del mismo modo, al contar con la misma escala la constante será cero.

$$\text{presente (p7A)} = 0 + 0{,}87_{\text{pasado (p7B)}} + e$$

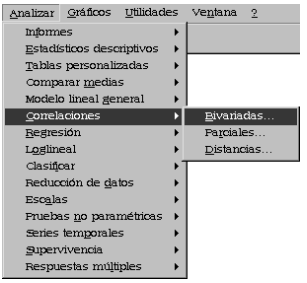


Figura 1

Una vez expuestos, desde un punto de vista teórico, los principales elementos que debemos considerar a la hora de abordar una análisis de regresión simple, su obtención informática parte de la consideración de tres cuadros de diálogos. Este proceso, como tenderemos ocasión de apreciar, se simplifica en el análisis de regresión múltiple.

4. Cuadro de Diálogo de Correlaciones Bivariadas

El primer paso a desarrollar consiste en determinar la efectiva y real relación lineal entre dos variables; esto es, debemos contar con la **matriz de correlaciones.** Para ello, en primer lugar, debemos elegir las dos variables que van a participar en la relación bivariada.

1er paso: Para poder seleccionar las dos variables seguiremos la secuencia **Analizar: Correlaciones: Bivariadas** (*figura 1*).

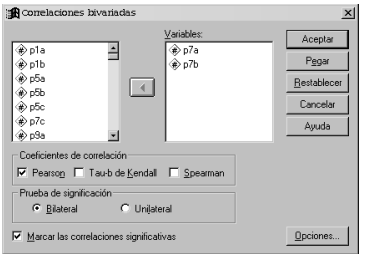


Figura 2

2º paso: Una ver en el Cuadro de Diálogo de correlaciones bivariadas seleccionaremos de la lista de variables las dos que nos interese relacionar. La selección deberá pasar al cuadro situado a la derecha (*figura 2*). En el ejemplo que reproducimos, hemos seleccionado las variables continuas p7A SITUACIÓN ACTUAL ESPAÑOLA (variable dependiente o criterio) y p7B SITUACIÓN ESPAÑOLA PASADA (variable independiente o predictora).Para valorar la relación y el ajuste de los datos al modelo de regresión debemos seleccionar, en el mismo cuadro de diálogo, el **Coeficiente de Correlación de Pearson y las Correlaciones significativas** con una **Prueba de Significación Bilateral.**

5. Cuadro de Diálogo de Gráficos de Dispersión

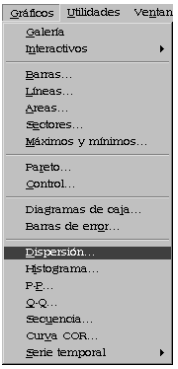


Figura 3

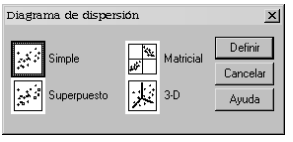


Figura 4

3er paso: Para valorar la bondad de ajuste de los datos podemos acompañar al coeficiente seleccionado de su correspondiente **Scatter Plot.** Dicho gráfico lo podemos seleccionar en **Gráficas: Dispersión: Simple** (*figuras 3 y 4*).

4º paso: Una vez en el cuadro de diálogo del Diagrama de dispersión simple (*figura 5*) debemos indicar la variable dependiente colocándola en el **Eje Y** así como la variable independiente situándola, en este caso, en el cuadro del **Eje X**. En este gráfico de dispersión de los valores X contra los de Y se observa la fuerza, dirección y forma de la relación entre las variables.

En el ejemplo que reproducimos, hemos seleccionado las variables continuas p7A SITUACIÓN ACTUAL ESPAÑOLA (variable dependiente o criterio) y p7B SITUACIÓN ESPAÑOLA PASADA (variable independiente o predictora). El gráfico de dispersión es el que aparece en el gráfico 1 que figura en el anexo de resultados. De su análisis podemos comprobar que en efecto existe relación entre las dos variables seleccionadas. La nube de puntos tiene una forma definida y los puntos se encuentran, más o menos, agrupados.

6. Cuadro de Diálogo del Análisis de Regresión Lineal

En la introducción del capítulo hemos presentado a los análisis de varianza y de residuales, como dos buenos criterios para valorar la bondad de ajuste de los datos al modelo de regresión

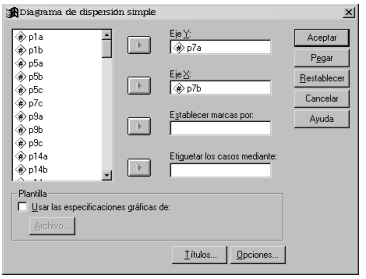


Figura 5

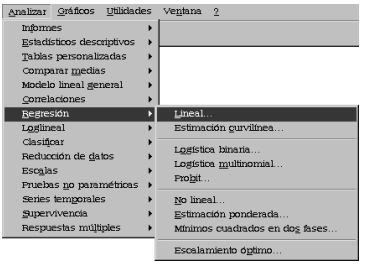


Figura 6

Figura 7

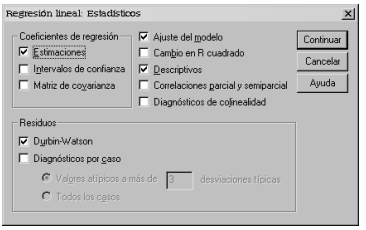


Figura 8

lineal. Estos análisis, así como los estadísticos R, R^2, R^2 Ajustada y Error Típico, solo se obtienen desde el Cuadro de Diálogo del Análisis de Regresión Lineal, el cual, se convierte en el cuadro específico del análisis de regresión múltiple al permitirnos seleccionar más de una variable independiente. A partir de ahora, y como ya hemos comentado, el Análisis de Regresión Bivariado, sigue el mismo procedimiento, como veremos más adelante, que el Análisis de regresión Múltiple. Lo único que va a cambiar es la elección del número de variables con las que vamos a trabajar. Por lo tanto, las salidas de ambos análisis será la misma.

Para completar la información respecto a la relación que se produce entre las dos variables debemos solicitarla, pues, en **el Cuadro de Diálogo correspondiente al Análisis de Regresión Lineal.**

5er paso: Al cuadro de diálogo se llega siguiendo la secuencia **Analizar: Regresión: Lineal** (*figura 6*).

6º paso: Una vez en él, deberemos especificar e introducir en sus correspondientes cuadros, la variable dependiente y la variable independiente (*figura 7*). Aquí nuevamente la variable p7A SITUACIÓN ACTUAL ESPAÑOLA hará las veces de variable dependiente o criterio y p7B SITUACIÓN ESPAÑOLA PASADA de variable independiente o explicativa.

7º paso: Cliqueando sobre el botón de comando **Estadísticas**, situado en la parte inferior del cuadro de diálogo principal, podremos *(figura 8*): obtener el estadístico **Durbin-Watson**, que nos permite realizar el análisis sobre los Residuos; los estadísticos **Descriptivos** básicos, que podremos utilizar en la interpretación y análisis de la relación (entre los que destaca la **Matriz de Correlaciones** para analizar la relación y significación); los estadísticos que nos permiten valorar el **Ajuste del modelo** (R, R^2, R^2 Corregida y el error típico de la estimación); por defecto, y también como criterio de validación del modelo nos presenta el análisis de varianza; y, por último, solicitaremos que nos calcule las **Estimaciones** de los **Coeficientes de regresión**, lo que nos permitirá concretar la ecuación predictiva.

8º paso: Como complemento al estadístico de Durbin-Watson cliqueando en el botón de **Gráficos** del cuadro de diálogo principal de regresión lineal podremos solicitar los gráficos ***Histograma y Gráfico de Probabilidad Normal*** (*figura 9*).

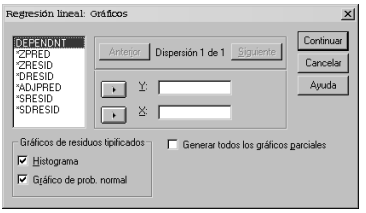


Figura 9

Los hechos y fenómenos sociales, complejos por naturaleza, son explicados no por una única causa sino por una pluralidad de ellas. Con la revolución informática se consolida la perspectiva del análisis multivariable relegando a un segundo plano el enfoque univariado del que el análisis de regresión simple forma parte. La aplicación de esta técnica, tal y como hoy es concebida y aplicada la investigación en ciencias sociales, se ciñe a dar respuesta puntual de alguna cuestión formando parte de una estrategia global de investigación. En este apartado recogemos cómo es aplicado este análisis en un aspecto muy puntual del análisis demográfico.

7. Bibliografía Comentada

• Díez Nicolás, Juan (1997): "La estructura de los hogares españoles", en Rafael Puyol (ed.) (1997), *Dinámica de la población en España. Cambios demográficos en el último cuarto del siglo XX*, Madrid, Síntesis, pp. 145-166.

El capítulo, entre otras muchas consideraciones, analiza la reducción del tamaño promedio de los hogares españoles. Con el objetivo de de comprobar si es la baja fecundidad o el alto número de hogares unipersonales el factor que más influye en el tamaño promedio de los hogares se ha aplicado un análisis de regresión simple tomando como unidad de análisis los 12 paises de la Unión Europea (este análisis también se ha aplicado para el conjunto de CCAA). Se ha comprobado que la correlación entre el índice de fecundidad y el tamaño medio de hogares es muy baja y negativa (r = -0.02); mientras que la correlación entre la proporción de hogares unipersonales sobre el total y el tamaño medio de los hogares es muy alta y también negativa (r= -0.90). El análisis ha demostrado que tanto para la Unión Europea como para el conjunto de CCAA, el tamaño medio de los hogares depende más de la proporción de hogares unipersonales que de la fecundidad en el sentido de que cuanto mayor el la proporción de hogares unipersonales sobre el total de hogares de una sociedad más pequeño es el tamaño medio de los hogares. Lo que explica la reducción del tamaño en los hogares en

la actualidad no es la baja fecundidad, sino el incremento en la proporción de hogares unipersonales.

8. Resultados

En el anexo que sigue se recogen las salidas de resultados que el paquete estadístico SPSS ofrece cuando es la técnica de regresión simple la que se ha aplicado.

- En primer lugar, el programa nos ofrece una tabla y un gráfico que dan cuenta de la bondad de ajuste de los datos al modelo de regresión simple: la tabla en donde aparece la **matriz de correlaciones** (cuadro de diálogo de correlación bivariada); y el **gráfico scatter plot** (del cuadro de diálogo de gráficas de dispersión).

A continuación se presentan el resto de tablas y gráficos obtenidos a partir de las restricciones impuestas en el cuadro de diálogo de regresión lineal (el que aplicaremos en el análisis de regresión múltiple).

- De este modo, se exponen las tablas que recogen información básica tanto del proceso como de las variables sometidas a análisis; esto es, la tabla de **descriptivos básicos** y la de **matriz de correlaciones parciales.**

- A continuación se presenta una tabla (resumen del modelo) en la que se relaciona una serie de estadísticos a partir de los cuáles valorar la **bondad de ajuste** de los datos del modelo. Con la misma finalidad se presenta la **tabla de análisis de varianza.** Incluye el estadístico Durbin-Watson que nos permite analizar la independencia de los residuales. Estas tablas, y los estadísticos adscritos a las mismas, complementan la información ya aportada en las primeras tablas de información.

- En tercer lugar, nos encontramos con la tabla en la que aparecen **los coeficientes** de la ecuación predictiva. Ésta se forma a partir de los coeficientes no estandarizados (B) cuando los valores de las dos variables tienen la misma escala. En el caso contrario deberemos elegir los coeficientes Beta.

- Por último, la exposición de resultados se cierra con dos representaciones gráficas cuya finalidad es facilitar el análisis respecto al tipo de distribución de los residuales: **gráfico de residuos tipificados y gráfico de probabilidad normal.**

8.1. Matriz de Correlaciones (Cuadro de diálogo de Correlaciones bivariadas)

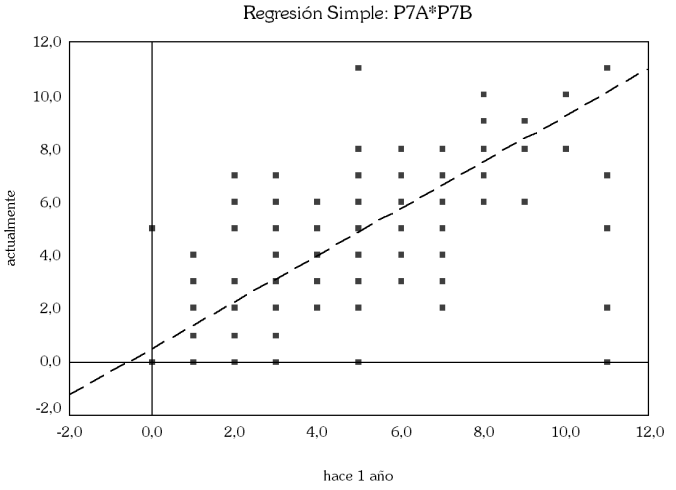
Correlaciones

		actualmente	hace 1 año
actualmente	**Correlación de Pearson**		
	Sig. (bilateral)		
	N		
hace 1 año	**Correlación de Pearson**	,871**	
	Sig. (bilateral)	,000	
	N	1200	

**. *La correlación es significativa al nivel 0,01 (bilateral).*

8.2. Gráfico de Dispersión (Cuadro de diálogo de Gráficos de Dispersión) (Gráfico 1).

Representación gráfica de la Regresión Simple



8.3. Estadísticos básicos (Cuadro de diálogo de Regresión Lineal)

Estadísticos descriptivos

	Media	Desviación típ.	N
actualmente	*4,79*	*2,14*	*1200*
hace 1 año	*4,91*	*2,13*	*1200*

8.4. Matriz de Correlaciones Parciales

Correlaciones

		actualmente	hace 1 año
Correlación de Pearson	**actualmente**	*1,000*	*,871*
	hace 1 año	*,871*	*1,000*
Sig. (unilateral)	**actualmente**	,	*,000*
	hace 1 año	*,000*	,
N	**actualmente**	*1200*	*1200*
	hace 1 año	*1200*	*1200*

8.5. Resumen del proceso STEPWISE: relación y eliminación de variables

Variables introducidas/eliminadas[a]

Modelo	Variables introducidas	Variables eliminadas	Método
1	*hace 1 año*	,	*Por pasos (criterio: Probabilidad de F para entrar <= ,050, Probabilidad de F para salir >= ,100).*

a. *Variable dependiente: actualmente*

8.6. Estadísticos de Bondad de Ajuste

Resumen del modelo[b]

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación	Durbin-Watson
1	*,871[a]*	*,759*	*,759*	*1,05*	*1,846*

a. *Variables predictoras: (Constante), P7B*

b. *Variable dependiente: P7A*

8.7. Tabla de Análisis de Varianza

ANOVA[b]

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	**Regresión**	*4148,999*	*1*	*4148,999*	*3773,649*	*,000[a]*
	Residual	*1317,160*	*1198*	*1,099*		
	Total	*5466,159*	*1199*			

a. *Variables predictoras: (Constante), hace 1 año*

b. *Variable dependiente: actualmente*

8.8. Estimaciones de parámetros o coeficientes de correlación: la ecuación de predicción

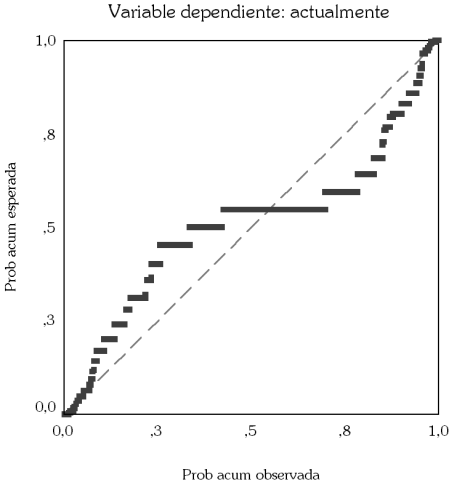
Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes estandarizados		
		B	Error típ.	Beta	t	Sig.
1	**(Constante)**	*,512*	*,076*		*6,738*	*,000*
	hace 1 año	*,872*	*,014*	*,871*	*61,430*	*,000*

a. *Variable dependiente: actualmente*

8.9. Grafico de distribución de residuales (gráfico 2)

Gráfico P-P normal de regresión Residuo tipificado



8.9. Grafico de probabilidad normal (gráfico 3)